STA-380 Exercises

Varun Kausika

2022-08-10

1. Probability practice

Part A

Given information:

Two categories of users:

1. Truthful clicker (TC)
2. Random clicker (RC)

Information on probablities:

- $P(RC) = 0.3$
- $P(Yes|RC) = 0.5$
- $P(No|RC) = 0.5$
- $P(TC) = 0.7$
- $P(Yes|TC) = x$
- $P(No|TC) = 1 - x$
- $P(Yes) = 0.65$
- $P(No) = 0.35$

Using the Rule of Total Probability,

$$P(Yes) = P(Yes, TC) + P(Yes, RC) = P(TC) * P(Yes|TC) + P(RC) * P(Yes|RC) \tag{1}$$

$$P(Yes) = 0.7x + 0.3 * 0.5 = 0.7x + 0.15 = 0.65$$

Solving for x, we get,

$x = P(Yes|TC) = 0.714$

Part B

We are being asked $P(Diseased|Positive)$

Given information:

- $P(Positive|Diseased) = 0.993$
- $P(Negative|NotDiseased) = 0.9999$
- $P(Diseased) = 0.000025$

According to Bayes Rule and Rule of Total Probability,

$$P(Diseased|Positive) = \frac{P(Positive|Diseased) * P(Diseased)}{P(Positive)} \tag{2}$$

and,

$$P(Positive) = P(Positive|Diseased) * P(Diseased) + P(Positive|Not\ Diseased) * P(Not\ Diseased) \tag{3}$$

Therefore,

$$P(Positive) = 0.993 * 0.000025 + 0.0001 * 0.999975 = 0.000125$$

Substituting in (2) we get,

$$P(Diseased|Positive) = \frac{0.993 * 0.000025}{0.000125} = 0.1986$$

2. Wrangling the Billboard Top 100

Part A

First, we load in the data and perform a group by on the performer and the song, with an agg function of count for the week

Table 1: Billboards

performer	song	count
'N Sync	(God Must Have Spent) A Little More Time On You	22
'N Sync	Bye Bye Bye	23
'N Sync	Gone	24
'N Sync	I Drive Myself Crazy	12
'N Sync	I Want You Back	24
'N Sync	It's Gonna Be Me	25
'N Sync	Pop	15
'N Sync	Tearin' Up My Heart	1
'N Sync	This I Promise You	26
'N Sync & Gloria Estefan	Music Of My Heart	20

Finally, we sort the dataframe in descending order of counts and find the top 10 and give our table a caption:

Table 2: Top 10 most popular songs

performer	song	count
Imagine Dragons	Radioactive	87
AWOLNATION	Sail	79
Jason Mraz	I'm Yours	76
The Weeknd	Blinding Lights	76
LeAnn Rimes	How Do I Live	69

performer	song	count
LMFAO Featuring Lauren Bennett & GoonRock	Party Rock Anthem	68
OneRepublic	Counting Stars	68
Adele	Rolling In The Deep	65
Jewel	Foolish Games/You Were Meant For Me	65
Carrie Underwood	Before He Cheats	64

Part B

First we group by year and order by ascending year. Then, we remove the years 1958 and 2021 from the rows and order just to make sure. Finally, we proceed to plot the columns.

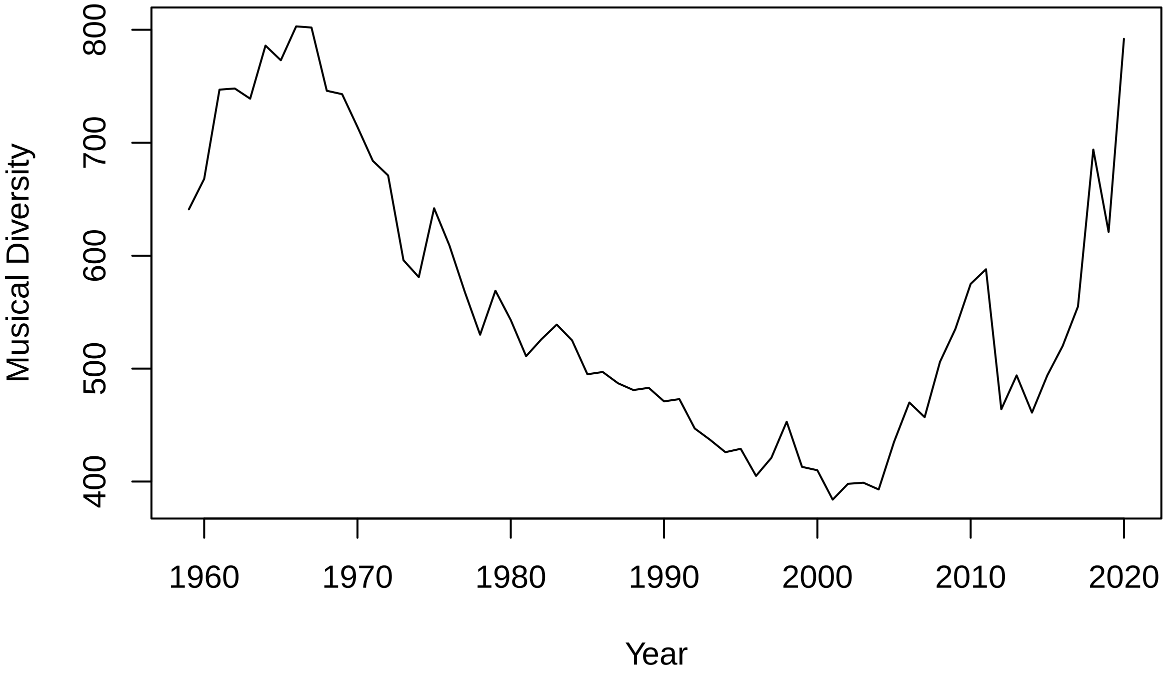

Figure 1: The plot shows peaks in diversity in 1965 and 2020, along with extreme lows in 2000

Part C

First, we filter the dataframe from part A to include only those songs with weeks at least 10. Then, We do a group by on the artists. Finally, we can select those artists with a hit-count of at least 30. **??**

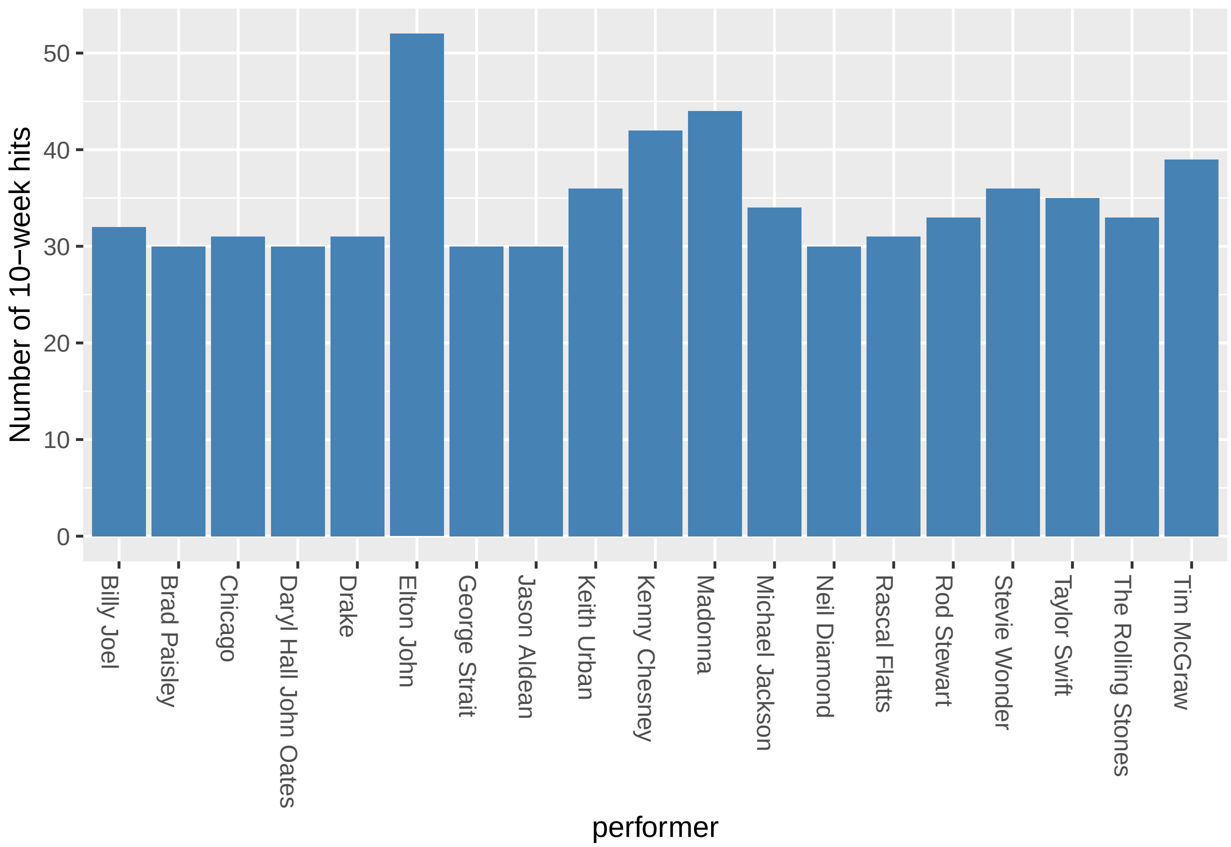

Figure 2: Elton John has more hits than others by quite a large margin

3. Visual Storytelling Part 1: green buildings

Outliers:

We are interested in finding the potential economic gains the owner could make by constructing a 15-story green building in the neighbourhood of East Cesar Chavez.

First, we remove the outliers from the dataset as mentioned by the analyst.

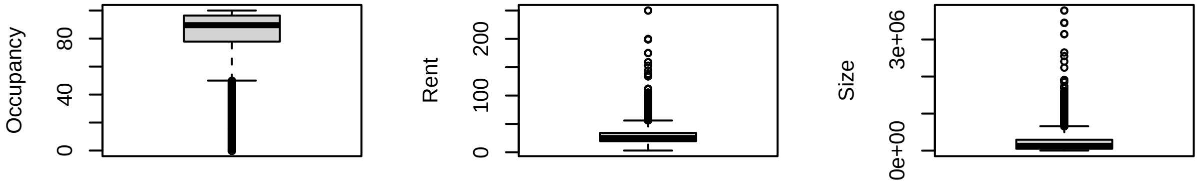

Figure 3: It seems like occupancy below 40 percent can be removed in contrast to the 10 percent that was suggested

Observations:

1. Removing outliers in occupancy below 40% would only remove 456 rows from our dataset, so we proceed with this step.
2. We should not remove outliers from rent, as that would shrink our dataset greatly, and also since most of the green buildings have higher rents, this would remove them entirely.
3. We should not remove outliers from size, because size and rent are positively correlated and this might remove a significant number of green buildings as well.

Correlation Matrix:

Now, we plot the feature correlation matrix in our dataset.

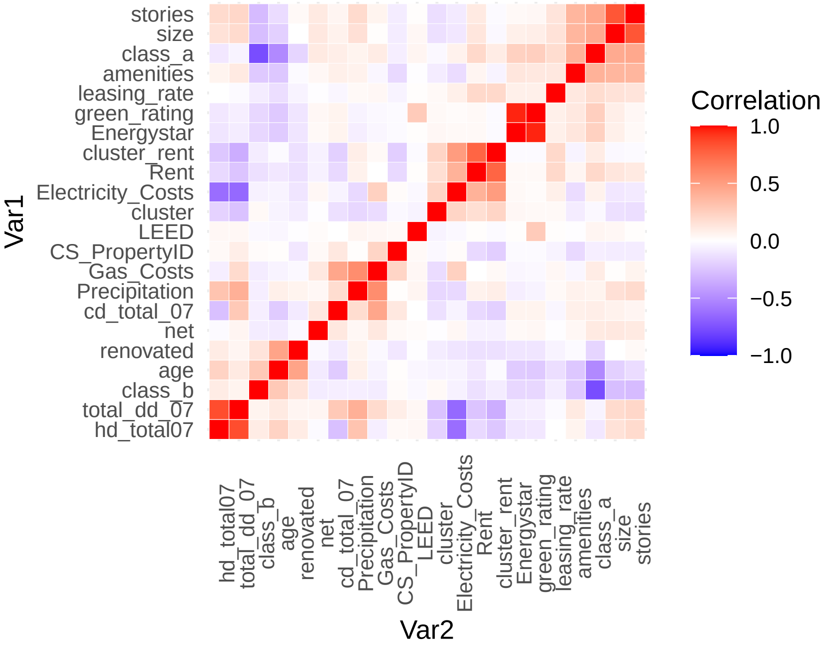

Figure 4: As we can see, rent is highly dependent on clusters

Observations:

1. Rent and Cluster has a high positive correlation, because some neighbourhoods are more expensive than others. So, instead of taking the median rent throughout our dataset, we can just use the cluster rent variable.
2. Rent and age has a mildly negative correlation, as one would expect.
3. Rent is not correlated with green rating, LEED or Energy Star.
4. Class A buildings are positively correlated with rent whereas Class B buildings are negatively correlated.
5. Leasing rate is positively correlated with rent.
6. Size and stories(can be considered correlated features) are both positively correlated to rent.

Scatter plots:

Now, we draw some scatter plots to understand the relation between correlated variables.

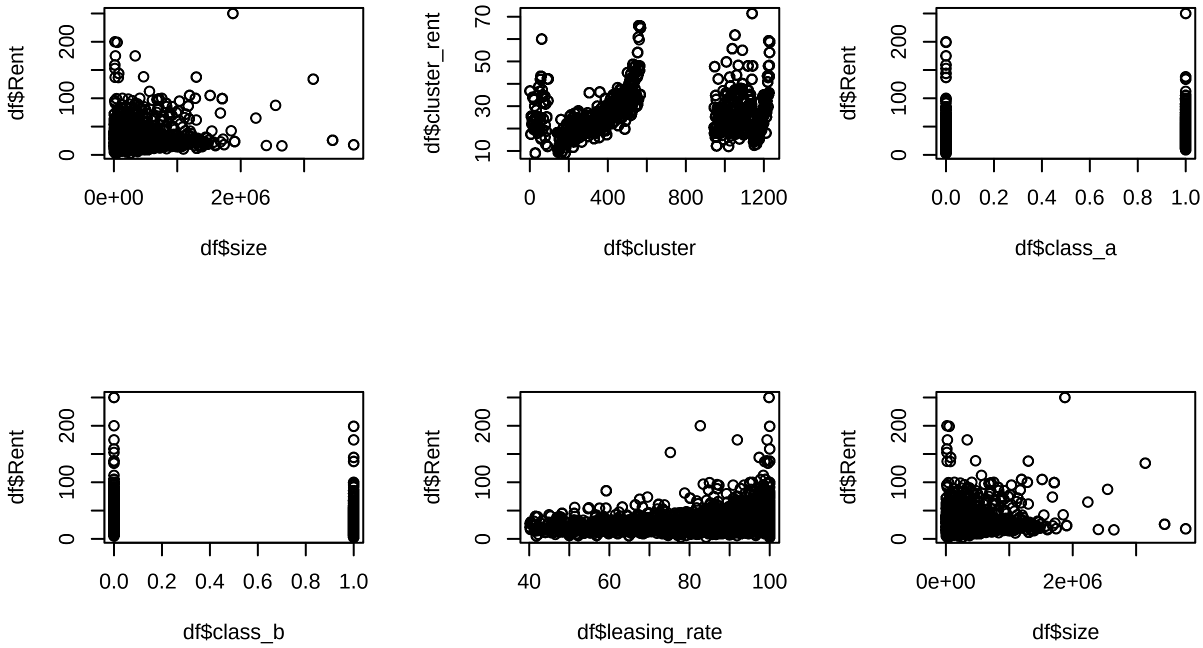

Figure 5: We can see clear patterns in the cluster rent v cluster. Some clusters have a significantly higher average rent.

Observations

1. Rent in Class A buildings occur more in positive class. This could be because of the positive correlation of class A buildings with green rating, size and stories.
2. Rent in Class B buildings occur more in negative class. The correlation of Class B with the variables is exactly opposite to that of Class A.
3. This implies that the builder **MUST** aim to construct a Class A building in order to be able to charge more rent.
4. Furthermore, the builder might want to move the building location to a cluster that falls in higher average rent bracket.

Density plots:

Now, we look at the green buildings in our dataset and compare them to all the other buildings

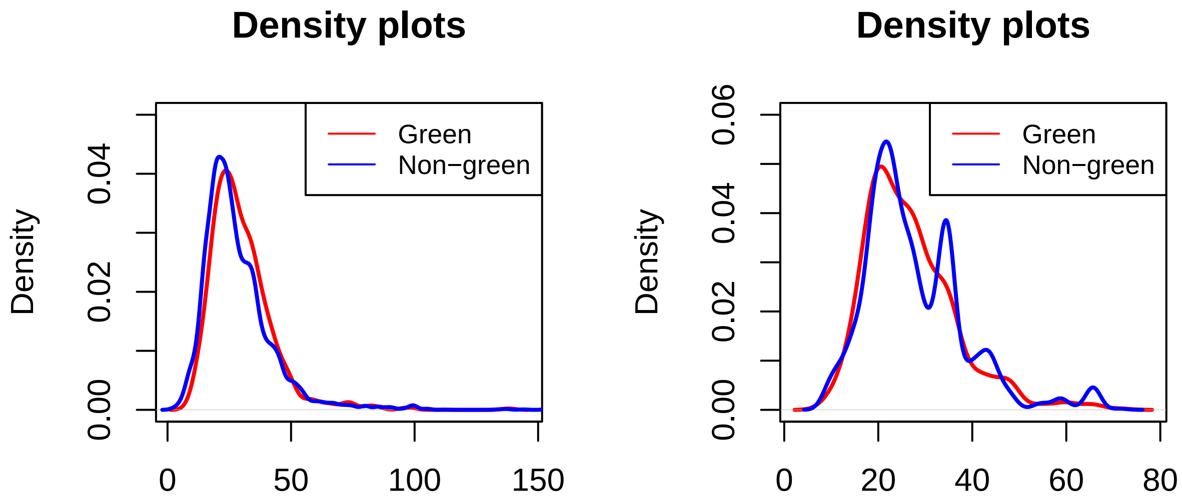

Figure 6: We can see that both green buildings and non-green buildings are not normally distributed. They are shifted to the left.

Observations:

1. We can conclude that this data is skewed toward the lower rent buildings (has long tails).
2. When comparing green and non-green buildings, the green buildings seem to peak at a slightly higher rent than the non green rents. Comparing the median values of both of these will be appropriate for the graph on the left.
3. However, for the cluster rent graph, we see that the peak of the non-green occurs at a slightly higher rent value than the peak for the green buildings.Therefore, if we use cluster rent as our primary variable to determine income (as we said we would before), we would expect in fact a lower rent value for the greens.
4. The cluster rent density function is not normal for either the greens or the non greens. For the non-greens, it looks more like a sum of many different bell shaped curves. This suggests that each cluster within the data has a separate normal distribution.

Economic Impact of Green Houses:

Economic impact could include a variety of factors in this case, namely:

1. **Sources of income depend on the rent being charged. Which inturn are caused by:**
   (a) Size of the house being rented (250,000 sqft in our case).
   (b) The neighbourhood of the house. A groupby statement is in order to find out the prices in East Cesar Chavez. On top of this, the premium rent for being green, as shown in the density plot, is pretty much no-existent. So controlling for the clusters, we can say that green houses don't do much better than non-green houses in this dataset.
   (c) Age of the building. One would expect that the older the building is the lower the rent is. This is shown in the correlation plot.

    (d) Appreciation of value of houses in the neighbourhood. This is something that cannot be measured by the given features. A thorough analysis of previous time series data of prices is recommended.
    (e) Occupancy of the building (relevant variable here is lease.rate).
    (f) Whether amenities are available or not.
    (g) Whether it was renovated or not.
    (h) Whether it is Class A or Class B.

2. **Sources of expenditure depend on:**
    (a) Initial capex (100,000 in our case).
    (b) Premium needed to be spent on constructing a green building (5%).
    (c) Maintainance, repairs and other charges.
    (d) **NOTE: Water, Gas and Electricity charges are assumed to be paid by the tenant, hence are not included.**

Where the Analyst went wrong:

The analyst found the impact on income in a very linear way. He did not take into account the following:

1. That the green buildings do not have any premium in rent after accounting for the cluster in which the buildings are.
2. That 40% of the occupancies are outliers, since they have a very high mean and low interquartile range.
3. Time value of money: he has calculated the simple Payback Period of the project. The more accurate estimator of success would have been the Net Present Value.
4. A cross sectional analysis across other buildings should have been done to compare Payback Periods instead of assuming 8 years is a decent time to recuperate costs
5. The worst case of occupancy is not 90%. It is much lower. As seen in the dataset many buildings have occupancies even lower than 40%.
6. That the cost of the buildings in the area may reduce/increase over time.
7. Accounting for other variables which may inhibit the ability to charge high rent (eg. Class A vs B, others listed above).

Confounding variables:

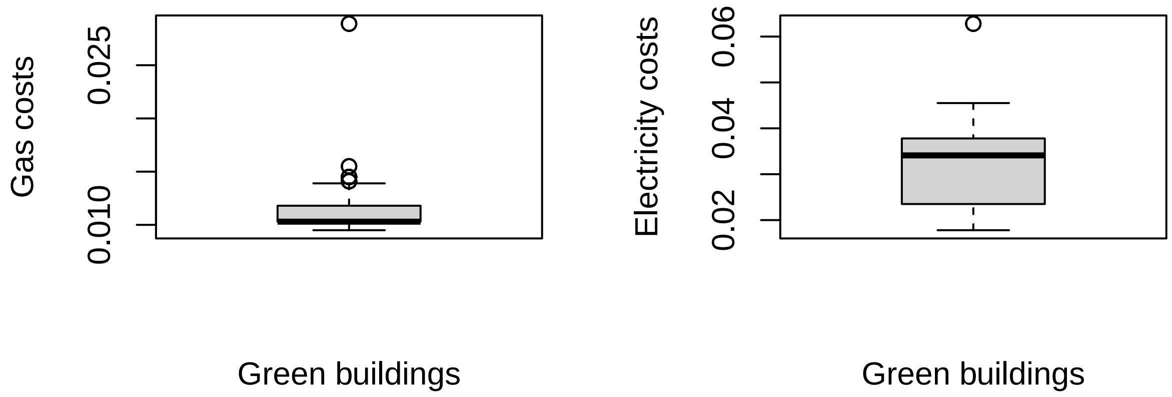

Figure 7: Some green houses show a high electricity and gas cost which goes against our intuition of what is green. The outliers can be adjusted for by removing them